Football Match Result Prediction - Project Protocol

Executive Summary

This project aims to predict football match outcomes?Home Win, Draw, or Away Win?using advanced match statistics and machine learning. The analysis covers the complete data science pipeline: from data preprocessing, exploratory data analysis (EDA), feature engineering, and selection, to classification model training and tuning. Among several tested algorithms, an XGBoost classifier emerged as the best performer, achieving high accuracy and generalization on unseen data.

1. Introduction

The dataset originates from Kaggle: https://www.kaggle.com/datasets/technika148/football-database. It aggregates comprehensive football match data from multiple European leagues and seasons. Each record represents a single match with features such as goals, assists, xG (expected goals), cards, and more. Predicting match outcomes using these features has practical applications in sports analytics, betting markets, and strategic coaching.

2. Objective

The main objective is to classify the outcome of football matches into three categories:

- Home Win

- Draw

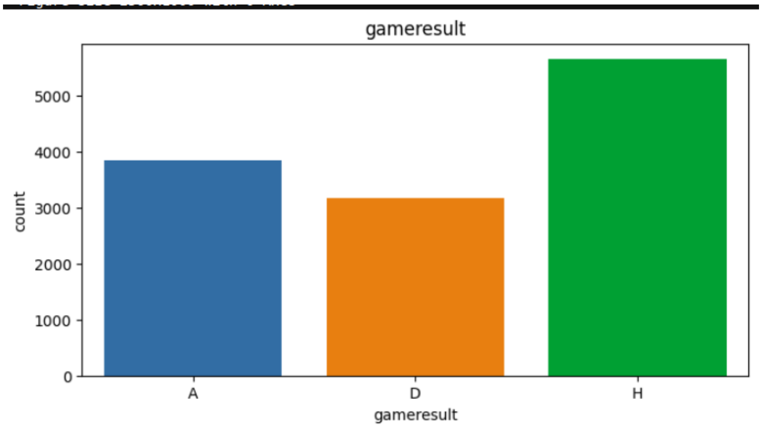
- Away Win

using historical and in-game statistics. The target variable is 'gameresult'.

3. Dataset Overview

Football Match Result Prediction - Project Protocol

The dataset includes 106 columns and over 7,000 match records. Variables cover team IDs, xG stats, cards, fouls, deep completions, and temporal information (season, date). Features are numeric, categorical, boolean, and datetime. A balance check shows relatively even distribution across the three classes.



4. Data Journey

Key steps included linking player and team IDs, creating time-based features (like month, day, weekend indicator), encoding categorical variables using ordinal encoders, and imputing missing values.

	playerID	teamID	playerName	teamName
0	560	89	Sergio Romero	Manchester United
1	557	89	Matteo Darmian	Manchester United
2	548	89	Daley Blind	Manchester United
3	628	89	Chris Smalling	Manchester United
4	1006	89	Luke Shaw	Manchester United
...	...	...	...	...
10101	7396	176	Loic Bessile	Bordeaux
10102	9566	175	Yanis Lhéry	Saint-Etienne
10103	9565	175	Mathys Saban	Saint-Etienne
10104	9568	181	Charles Costes	Dijon
10105	9567	181	Erwan Belhadji	Dijon

5. Methodology

The pipeline included several stages:

a) Preprocessing: cleaned and converted data types, removed irrelevant columns.

b) EDA: investigated distribution, outliers, and class balance.

c) Feature Engineering: created ratio features, rolling averages, categorical bins.

d) Feature Selection: used Lasso, Random Forest, XGBoost importances.

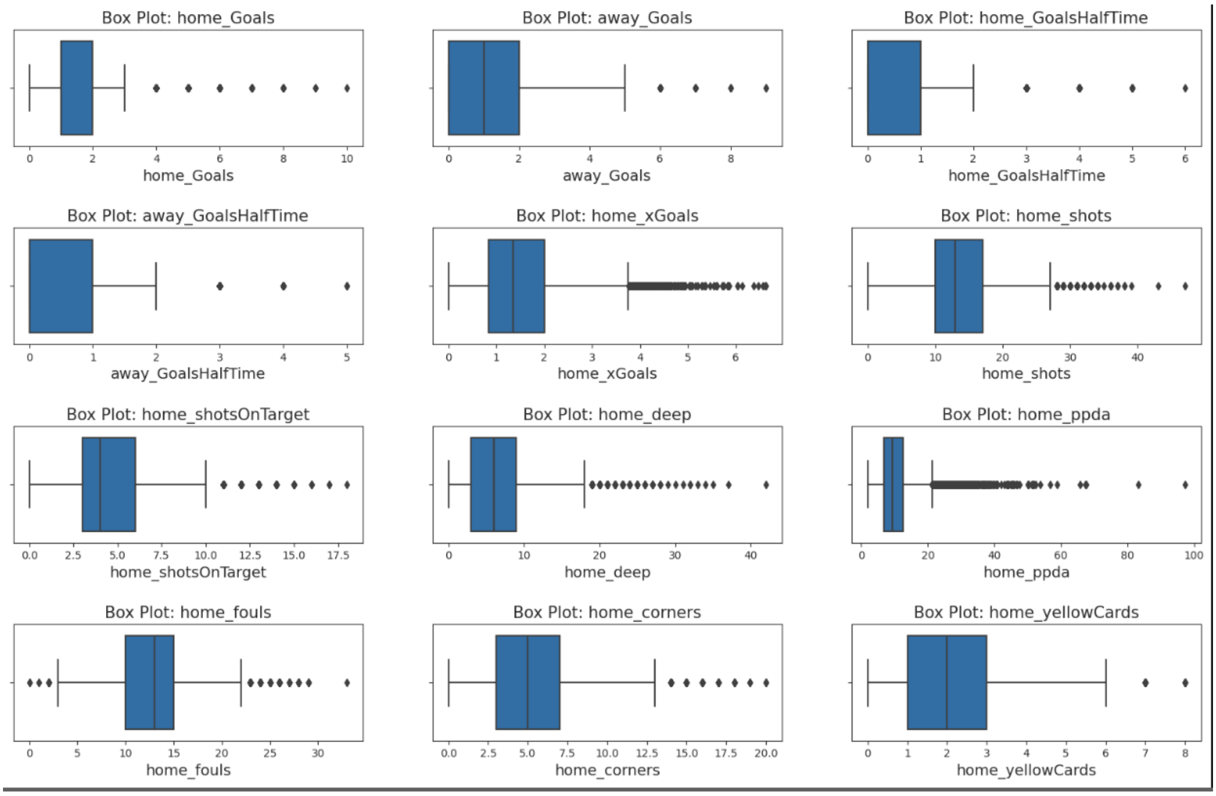
e) Modeling: trained multiple classifiers and tuned hyperparameters.

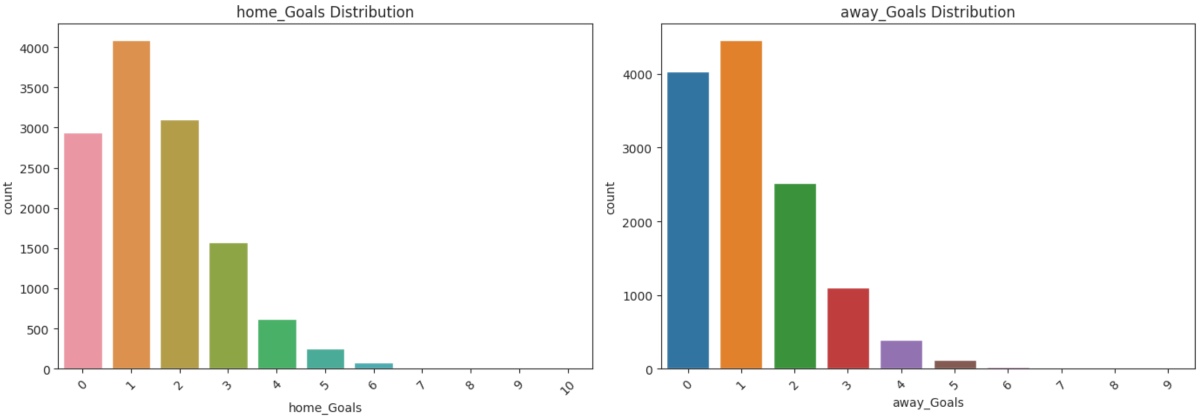
5.1 Exploratory Data Analysis (EDA)

Football Match Result Prediction - Project Protocol

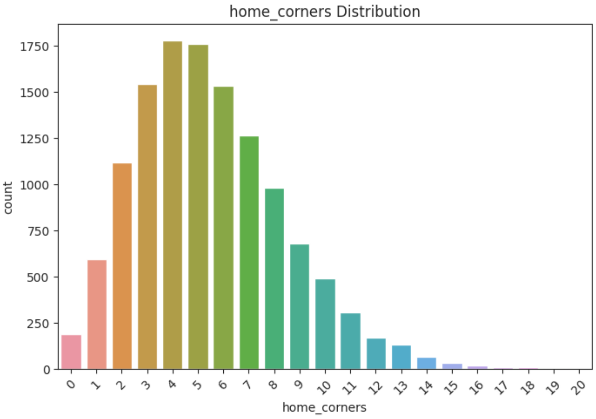
T-tests and Chi-Square tests were used to identify statistically significant features across match outcomes.

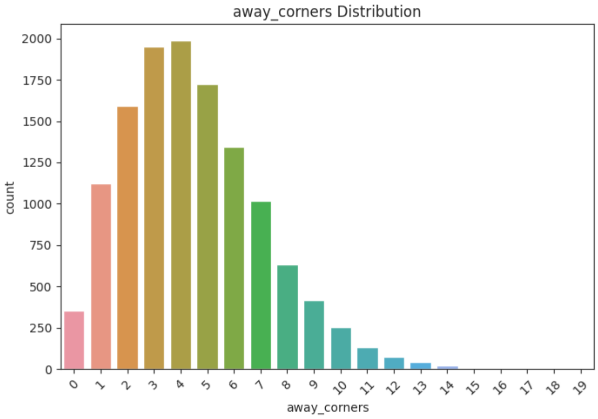
Visual tools like histograms, countplots, and boxplots supported this process.

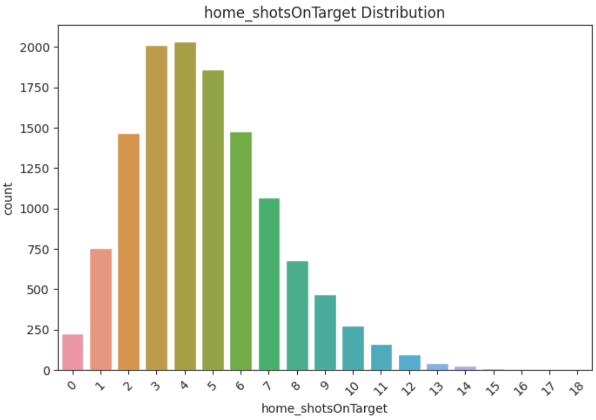


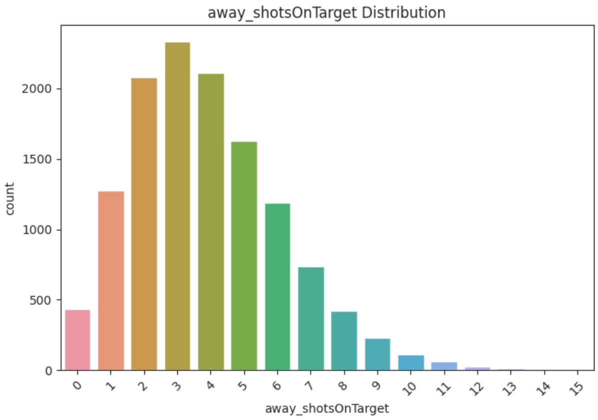


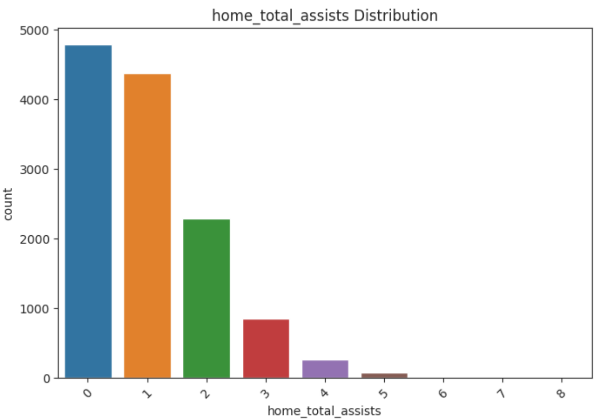
Football Match Result Prediction - Project Protocol

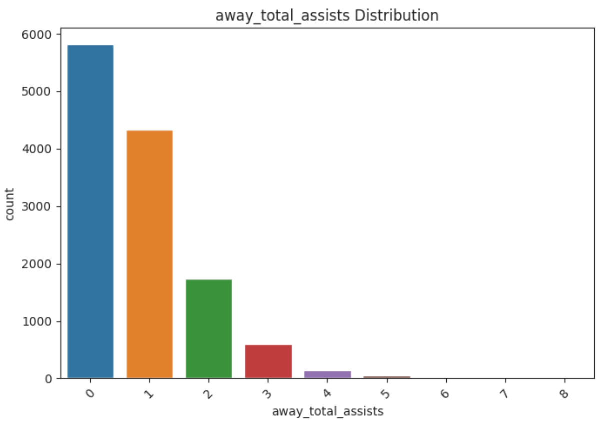


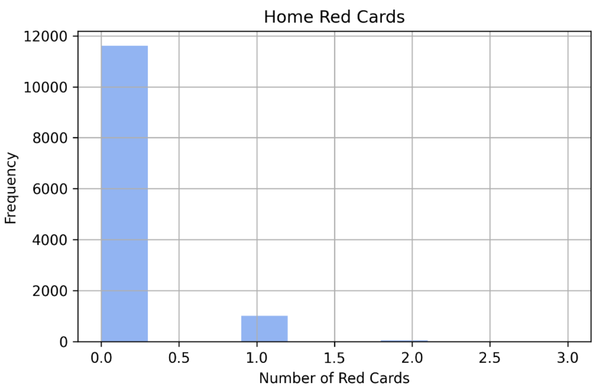


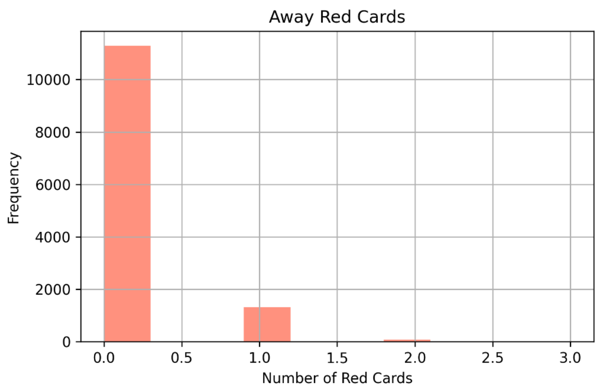




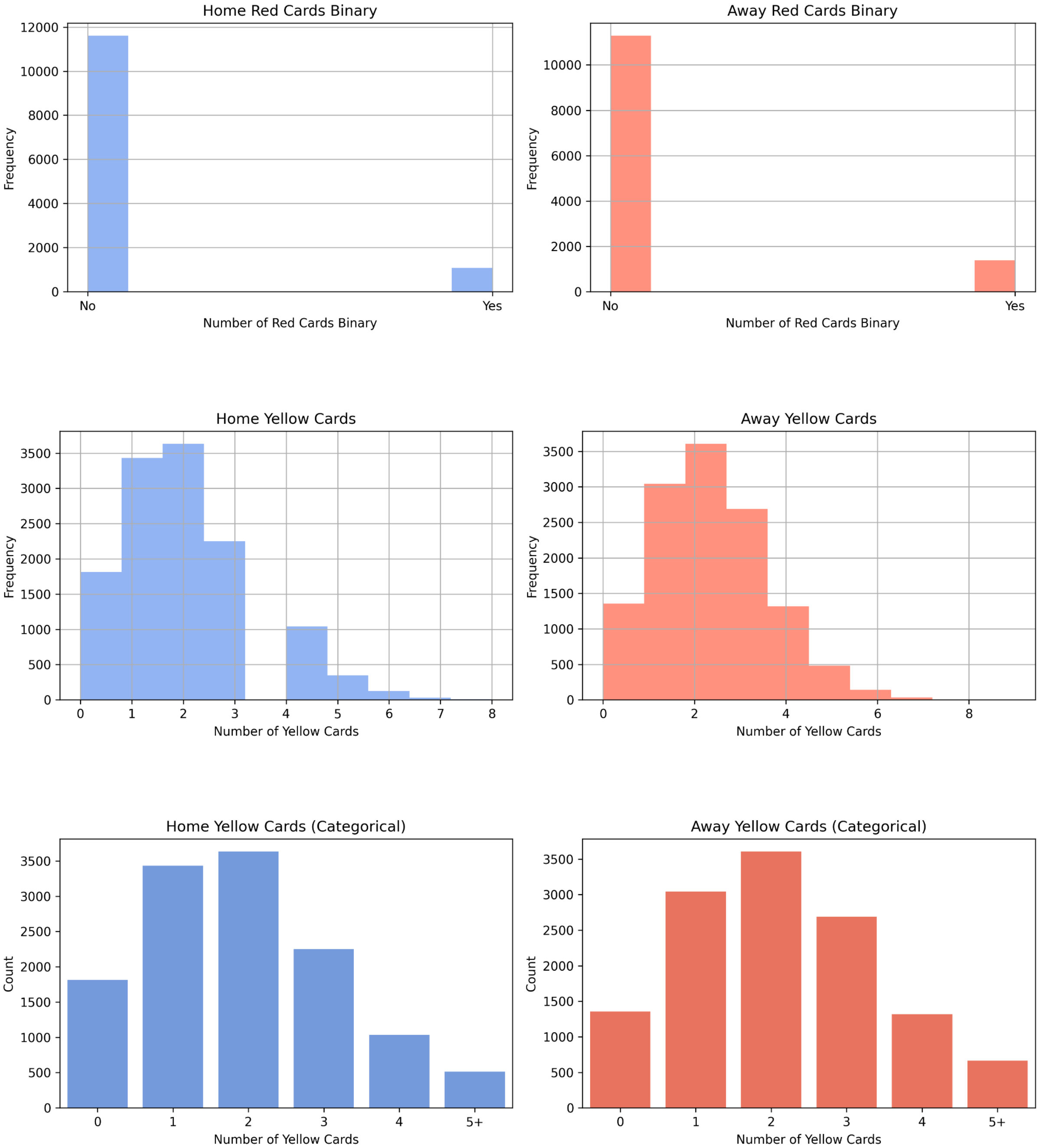




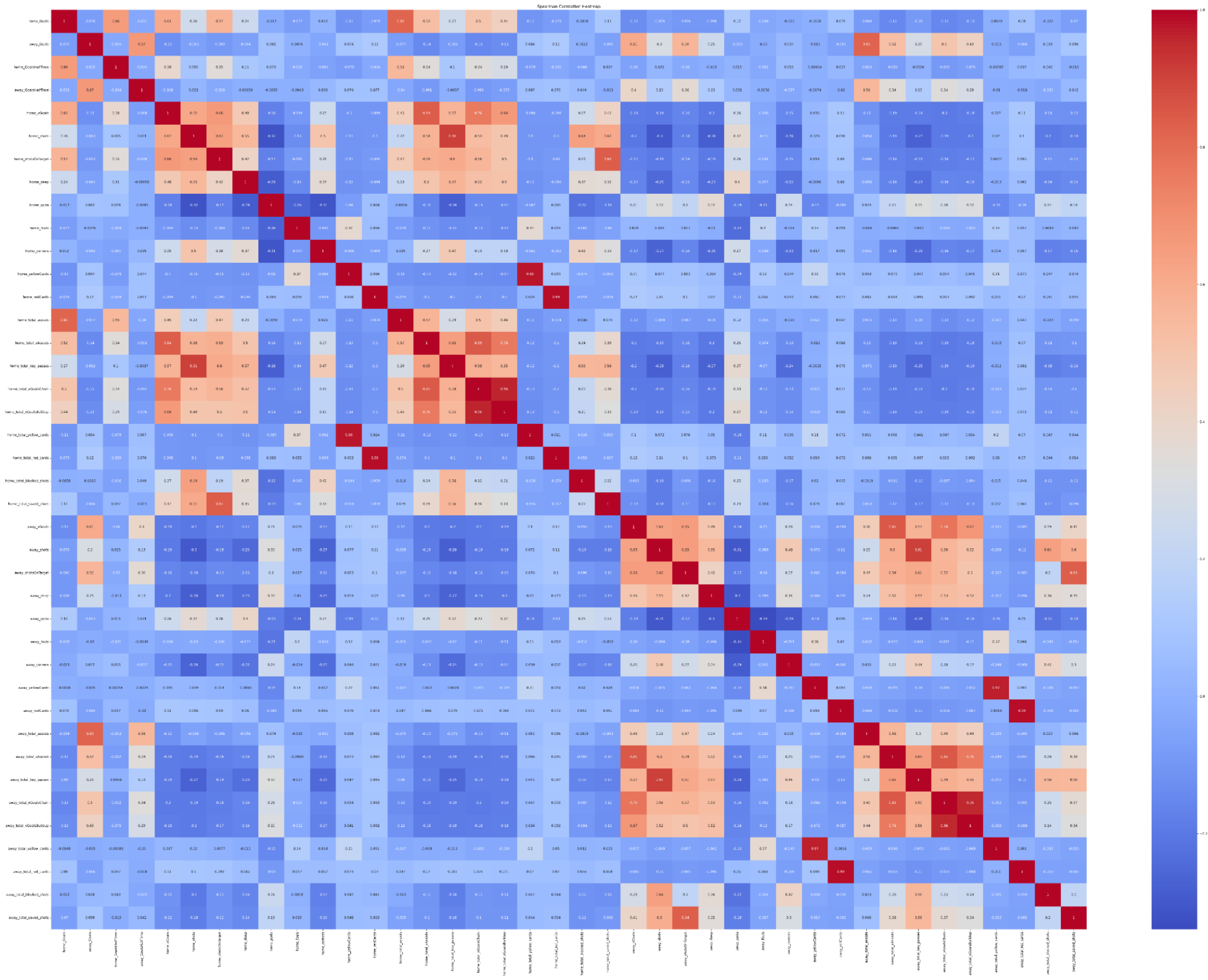




Football Match Result Prediction - Project Protocol

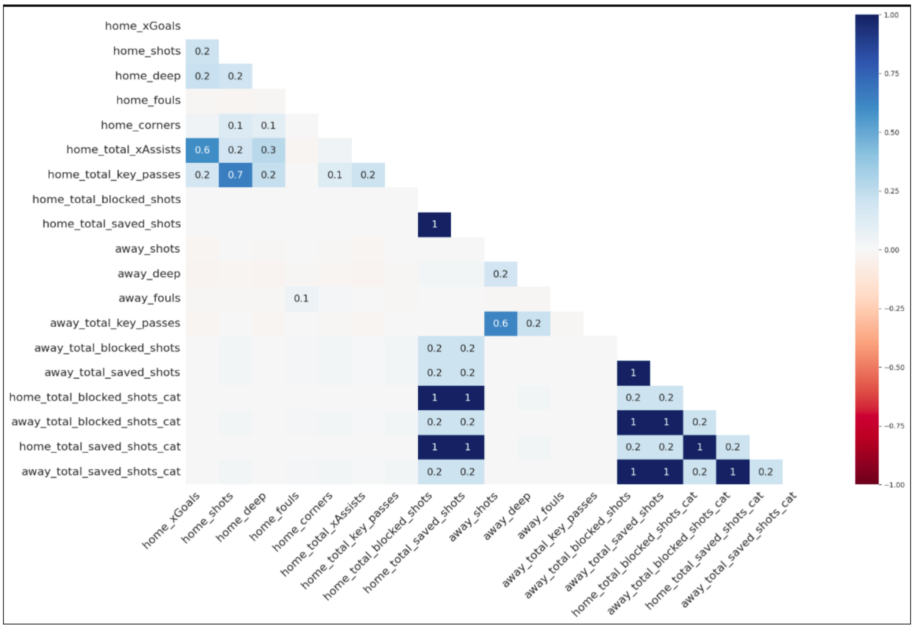


Football Match Result Prediction - Project Protocol



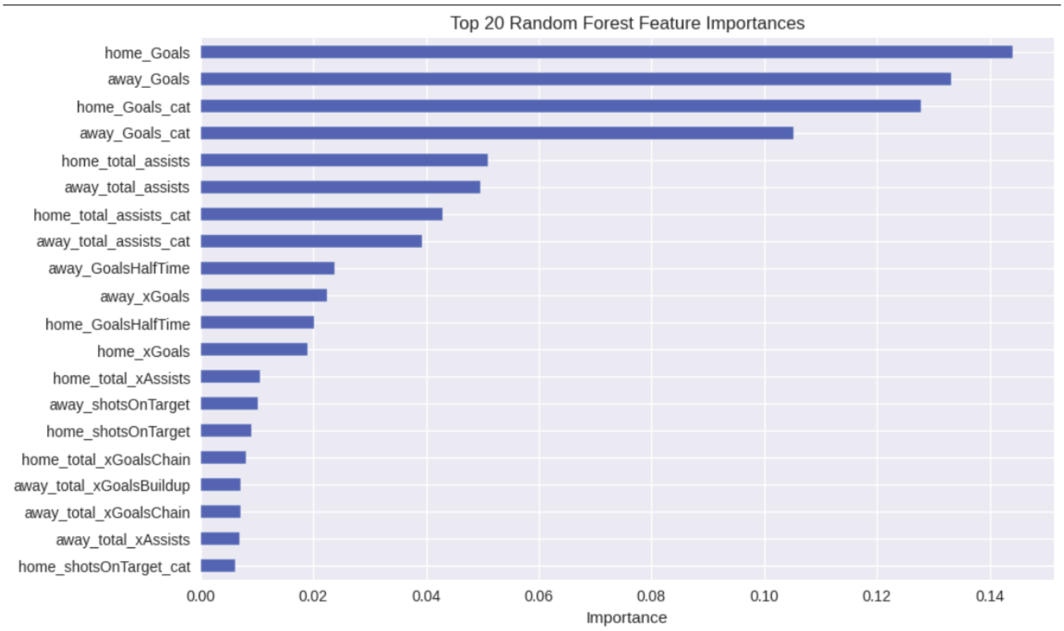
5.2 Handling Missing Values

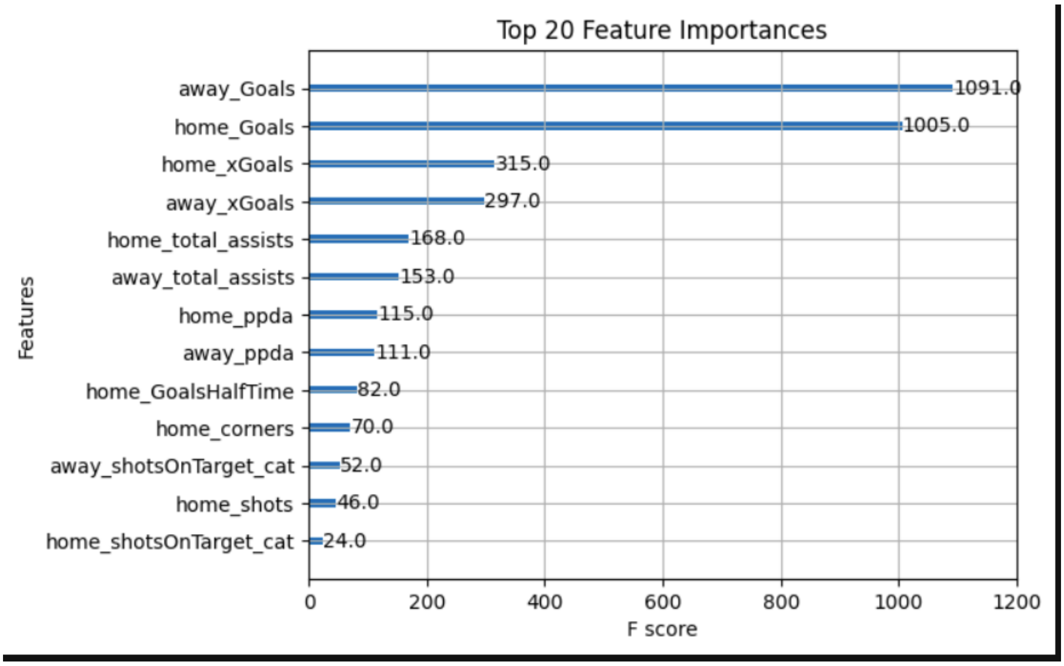
Missing values were visualized and imputed. For numerical columns, mean or forward fill was used. Categorical values were filled using the mode or converted to new 'unknown' categories.



5.3 Feature Engineering & Selection

Derived features included goals per shot, blocked shot ratios, and form vs discipline scores. Rolling averages and interaction terms between teams were also computed. Selection was done using statistical tests, Lasso regularization, and model-based importance rankings (RF and XGBoost).





6. Model Training & Evaluation

A variety of classifiers were trained using stratified train-test splits. Hyperparameter tuning was performed using RandomizedSearchCV and GridSearchCV. XGBoost outperformed others with high accuracy and low log-loss on the test set.

Model	**Accuracy**	**F1**	**LogLoss**	**AUC**
Logistic Regression	1.000	1.000	0.0052	1.000
XGBoost	0.998	0.998	0.0052	1.000
GBM	0.999	0.999	0.0072	0.9999
Random Forest	0.998	0.998	0.0150	0.9999
SVM	0.994	0.994	0.0122	0.9999
Decision Tree	0.998	0.998	0.0710	0.9981
AdaBoost	0.662	0.681	0.6453	0.9631

7. Final Model Deployment

The XGBoost model (max_depth=110, learning_rate=0.05, n_estimators=400) showed stable results across all data splits. It can be integrated into an application pipeline for match prediction, betting odds estimation, or analytics dashboards.

8. Conclusion

The football prediction project demonstrated a complete machine learning workflow with practical results. Insights from statistical testing and feature engineering significantly improved model performance. Future work may include using sequence modeling for player-level time-series or adding live-match data.

Thank you,

Leonardo Romano